Protein Function Prediction using Multi-label Ensemble Classification

supplementary file

Guoxian Yu, Huzefa Rangwala, Carlotta Domeniconi, Guoji Zhang, and Zhiwen Yu, *Member, IEEE*

This is a supplementary file of "Protein Function Prediction using Multi-label Ensemble Classification", which will be published by *IEEE Transactions on Computational Biology and Bioinformatics*.

1 Evaluation Metrics

This section appends the introduction of evaluation metrics (Section 4.2).

Ranking loss evaluates the average fraction of function label pairs that are not correctly ordered:

$$RankingLoss = \frac{1}{u}\sum_{i=l+1}^{N}\frac{1}{|\mathcal{Y}_i||\bar{\mathcal{Y}}_i|}|\{(c_1,c_2)\in \mathcal{Y}_i\times\bar{\mathcal{Y}}_i|F(i,c_1)\leq F(i,c_2)\}|$$

where $\mathcal{Y}_i$ is the function set of protein i, and $\bar{\mathcal{Y}}_i$ is the complement set of $\mathcal{Y}_i$. The performance is perfect when *RankingLoss=0*, and the smaller the value, the better the performance.

Coverage evaluates how far, on average, we need to go down the label ranking list to cover all the ground-truth labels of the instance:

$$Coverage = \frac{1}{u}\sum_{i=l+1}^{N}\max_{c\in\mathcal{Y}_i} rank(F(i,c)) - 1$$

Coverage is often bigger than 1, and the lower the value, the better the performance.

G. Yu is with the College of Computer and Information Science, Southwest University, Chongqing, 410075 China, and School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510006 China, email: gxyu@swu.edu.cn

H. Rangwala and C. Domeniconi are with the Department of Computer Science, George Mason University, Fairfax, VA, 22030 USA, email: rangwala@cs.gmu.edu, carlotta@cs.gmu.edu

G. Zhang is with the School of Sciences, South China University of Technology, Guangzhou, 510640 China, email: magjzh@scut.edu.cn

Z Yu is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510006 China, email: zhwyu@scut.edu.cn

Manuscript received 4 Apr. 2013; revised 17 Aug. 2013; accepted 29 Aug. 2013; published online xx xxx. 2013

2 Influence of Different Weighting Schemes

This section appends the results of ***Coverage*** *(in Table 1) with respect to different weighting scheme (Section 5.3).*

3 Influence of Different Ensemble Techniques

This section appends the introduction on different ensemble techniques and the experimental results (Table 2) with respect to Coverage *(Section 5.4).*

Decision templates uses the similarity between the matrix of classifier outputs for an input $\mathbf{x}$ (the Decision Profile $DP(\mathbf{x})$) and the C matrix templates found as the class means of the classifier outputs to determine the likelihood of $\mathbf{x}$ with respect to a specific class [1], [2]. $DP(\mathbf{x})$ for a sample $\mathbf{x}$ is a matrix composed of the entries $d_{r,c}(\mathbf{x})\in[0,1]$ $(r=1,...,R; c=1,...,C)$, each entry represents the support given by the r-th classifier to class c. Decision templates DT_c are the averaged decision profiles obtained from $\mathbf{X}_c$, which includes the set of training samples belonging to the c-th class.

$$DT_c = \frac{\sum_{\mathbf{x}\in\mathbf{X}_c} DP(\mathbf{x})}{|\mathbf{X}_c|}$$

For a test sample $\mathbf{x}$, its similarity between decision profile $DP(\mathbf{x})$ and the c-th decision template DT_c is computed as:

$$sim(\mathbf{x},c) = 1 - \frac{1}{R\times C}\sum_{r=1}^{R}\sum_{m=1}^{C}(DT_c(r,m) - d_{r,m}(\mathbf{x}))^2$$

$sim(\mathbf{x},c)$ can be viewed as the predicted likelihood of $\mathbf{x}$ with respect to the c-th class.

The linear regression based ensemble combines R classifiers by optimizing a linear regression problem as follows [3]:

$$\begin{aligned}\arg\min_{\boldsymbol{\omega}_c}\sum_{r=1}^{R}\sum_{i=1}^{l}(\omega_{rc}F_r(\mathbf{x}_i,c)-\mathbf{y}_{ic})^2+\mu\|\boldsymbol{\omega}_c\|^2 \\ = \arg\min_{\boldsymbol{\omega}_c}\|\mathbf{F}_{lc}\boldsymbol{\omega}_c-\mathbf{Y}_{lc}\|^2+\mu\|\boldsymbol{\omega}_c\|^2\end{aligned}$$

TABLE 1
Coverage (avg±std) on **Yeast (44 kernels)**, **Human (8 kernels)** and **Fly (38 kernels)**, with respect to different weighting schemes.

	Yeast		Human		Fly	
	20%	50%	20%	50%	20%	50%
TMC-B	25.16±0.67	22.04±0.57	84.74±1.98	65.23±1.69	188.79±3.35	153.30±2.45
TMC-BW	23.25±0.65	20.06±0.51	83.78±1.97	63.99±1.65	178.35±3.23	141.71±2.53
TMC-E	22.61±0.67	18.81±0.44	93.32±2.07	78.73±1.47	173.48±2.86	135.80±2.83
TMEC	**21.30±0.52**	**17.72±0.40**	**79.07±1.60**	**59.67±1.55**	**166.80±2.39**	**132.02±3.19**

where ω_{rc} is the weight of the r-th classifier with respect to the c-th label. $\mathbf{F}_c = [\mathbf{F}_{1c}, \mathbf{F}_{2c}, \cdots, \mathbf{F}_{Rc}] \in \mathbb{R}^{N\times R}$, $\mathbf{F}_{rc} \in \mathbb{R}^N$ is F_r predicted likelihood vector with respect to the c-th label on N samples, and $\mathbf{F}_{lc}$ is the first l rows of $\mathbf{F}_c$. $\mathbf{y}_{ic} = 1$ if the i-th sample has the c-label, and $\mathbf{y}_{ic} = 0$ otherwise. $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N] \in \mathbb{R}^{N\times C}$ is the initial labels on N samples and $\mathbf{Y}_{lc}$ is the first l rows with respect to the c-th column of $\mathbf{Y}$. μ is set to balance the importance of the first term and the second term. Our aim in using ω_{rc} is that different classifiers may have different weights on the c-th label. The analytical solution of $\boldsymbol{\omega}_c$ is:

$$\boldsymbol{\omega}_c = (\mathbf{F}_{lc}^{\mathrm{T}}\mathbf{F}_{lc} + \mu\mathbf{I}_R)^{-1}\mathbf{F}_{lc}^{T}\mathbf{Y}_{lc}$$

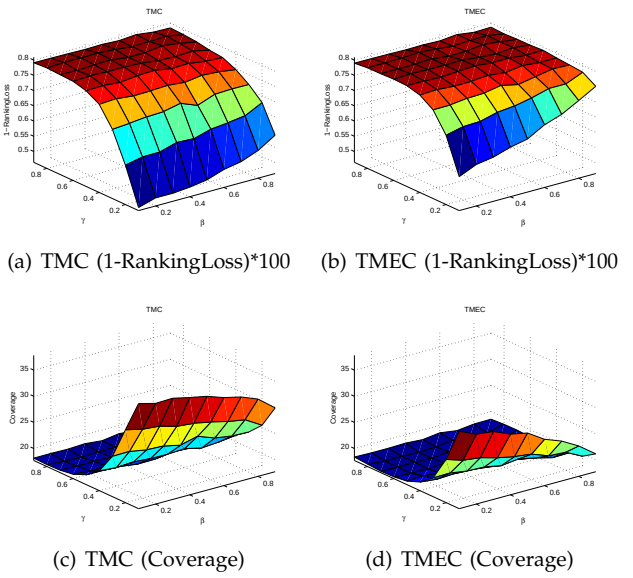
where $\mathbf{I}_R$ is a $R \times R$ identity matrix. Let $\mathbf{W} = [\boldsymbol{\omega}_1^{\mathrm{T}}, \boldsymbol{\omega}_2^{\mathrm{T}}, \cdots, \boldsymbol{\omega}_C^{\mathrm{T}}] \in \mathbb{R}^{R\times C}$, then the ensemble predicted likelihood vector is:

$$f(\mathbf{x}) = \sum_{r=1}^{R} \mathbf{W}_r \circ F_r(\mathbf{x})$$

where $\circ$ is the element-wise multiplication operator.

TABLE 2
Performance (*Coverage*) with respect to different ensemble techniques.

Methods	Yeast	Human	Fly
TMEC	**18.20±0.54**	**62.99±1.55**	**145.45±3.73**
TMEC-DT	27.28±0.25	111.82±0.76	189.87±1.52
TMEC-Reg	30.97±0.73	115.05±3.42	254.71±6.23



(a) TMC (1-RankingLoss)*100 (b) TMEC (1-RankingLoss)*100

(c) TMC (Coverage) (d) TMEC (Coverage)

Fig. 1. *1-RankingLoss* and *Coverage* on different β and γ **(Yeast)**. Similar colors in the figure are parameters with similar predictive performance.

4 Parameter Sensitivity Analysis

*This section appends the results of the parameter sensitivity analysis (Fig. 1) on **Yeast** (Section 5.5).*

References

[1] L. I. Kuncheva, J. C. Bezdek, and R. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," *Pattern Recognition*, vol. 34, no. 2, pp. 299–314, 2001.

[2] M. Re and G. Valentini, "Ensemble based data fusion for gene function prediction," *Multiple Classifier Systems*, pp. 448–457, 2009.

[3] G. Yu, G. Zhang, Z. Zhang, Z. Yu, and L. Deng, "Semi-supervised classification based on subspace sparse representation," *Knowledge and Information Systems, Minor Revision and Resubmitted on 2013-06-29.*